Computational Screening of Bacterial Natural Products for Antimicrobial Discovery: An Unsupervised Learning Approach with Bimodal Selection Strategy

K-Dense Web

Computational Drug Discovery Unit

correspondence@kdenseweb.org

December 2025

Abstract

Background: Antimicrobial resistance (AMR) represents one of the most critical global health challenges, with bacterial AMR directly attributable to over 1.14 million deaths annually. Natural products, particularly those derived from bacterial sources such as actinomycetes, have historically been the primary source of antibiotic scaffolds. However, systematic computational approaches for prioritizing natural product libraries for antimicrobial screening remain underexplored.

Methods: We developed a comprehensive computational pipeline to screen the COCONUT natural products database for potential antimicrobial candidates. Starting with 715,822 compounds, we filtered for bacterial-derived natural products (24,910 compounds), computed molecular descriptors using RDKit, and applied unsupervised machine learning including Principal Component Analysis (PCA), K-Means clustering, UMAP visualization, and DBSCAN outlier detection to characterize the chemical space. A bimodal selection strategy was employed to prioritize candidates balancing drug-likeness and structural complexity.

Results: Chemical space analysis revealed two distinct clusters: Cluster 0 (75.4%, 18,780 compounds) comprising smaller, drug-like molecules (mean MW: 396 Da, mean QED: 0.44), and Cluster 1 (24.6%, 6,130 compounds) containing larger, structurally complex natural products (mean MW: 978 Da, mean QED: 0.09). The bimodal selection strategy identified 50 prioritized candidates: Group A (n=25) drug-like leads with high QED scores (mean: 0.93) and Lipinski compliance, and Group B (n=25) complex scaffolds with privileged antibiotic architectures (mean MW: 1,930 Da).

Conclusions: Our unsupervised learning approach successfully characterized bacterial natural product chemical space and identified 50 high-priority candidates for experimental antimicrobial screening. The bimodal strategy ensures coverage of both developable leads and structurally novel scaffolds, maximizing the probability of identifying compounds with potent and novel mechanisms of action against drug-resistant bacteria.

Keywords: Natural products, antimicrobial resistance, drug discovery, machine learning, chemical space, clustering, COCONUT database

1 Introduction

1.1 The Antimicrobial Resistance Crisis

Antimicrobial resistance (AMR) has emerged as one of the most pressing global health threats of the 21st century. According to the Global Burden of Disease study, bacterial AMR was directly responsible for 1.14 million deaths in 2021 and associated with 4.71 million deaths worldwide [GBD 2021 Antimicrobial Resistance Collaborators, 2024]. This mortality burden exceeds that of HIV/AIDS and malaria combined, establishing AMR as a leading cause of death globally [Antimicrobial Resistance Collaborators, 2022]. The World Health Organization's 2024 Bacterial Priority Pathogens List identifies 15 resistant bacterial families requiring urgent attention, with projections indicating a twofold rise in resistance to last-resort antibiotics by 2035 [World Health Organization, 2024].

The crisis is particularly acute for multidrug-resistant (MDR) organisms including methicillin-resistant *Staphylococcus aureus* (MRSA), carbapenem-resistant *Enterobacteriaceae* (CRE), and MDR *Pseudomonas aeruginosa.* Recent meta-analyses report MDR prevalence rates exceeding 46% for *P. aeruginosa* isolates in Asia and Africa [Karruli et al., 2025]. Without intervention, AMR-attributable deaths are projected to reach 1.91 million annually by 2050, a 69.6% increase from 2022 levels [GBD 2021 Antimicrobial Resistance Collaborators, 2024].

1.2 Natural Products as Antibiotic Sources

Natural products have historically been the cornerstone of antibiotic discovery, with approximately two-thirds of approved antibiotics derived from microbial secondary metabolites [Newman and Cragg, 2020, Atanasov et al., 2021]. Actinomycetes, particularly *Streptomyces* species, represent the most prolific source of antibacterial natural products, producing diverse polyketides, non-ribosomal peptides, and terpenoids through specialized biosynthetic gene clusters [Sun et al., 2025, Editorial Board, 2024].

Recent advances in genome mining have revealed that actinomycete genomes harbor substantial untapped biosynthetic potential, with an estimated 70% of biosynthetic gene clusters remaining cryptic under standard laboratory conditions [Chen et al., 2021]. Marine-associated actinomycetes have emerged as particularly promising sources, with compounds such as isoquinocycline B demonstrating potent activity against *Bacillus subtilis* (MIC <0.048 μM) [Girão et al., 2021].

1.3 Computational Approaches to Natural Product Screening

The COCONUT database represents the most comprehensive open resource for natural product structures, aggregating over 695,000 compounds from 63 sources with annotations including organism origin and biosynthetic pathway classifications [Chandrasekhar et al., 2025]. This resource enables systematic computational screening approaches that can dramatically reduce the experimental burden of natural product drug discovery.

Modern cheminformatics tools, particularly RDKit, enable rapid calculation of molecular descriptors essential for evaluating drug-likeness [Landrum, 2024]. Key metrics include Lipinski's Rule of Five parameters (molecular weight, LogP, hydrogen bond donors and acceptors) [Lipinski et al., 2001], the Quantitative Estimate of Drug-likeness (QED) [Bickerton et al., 2012], and structural alerts for pan-assay interference compounds (PAINS) [Baell and Holloway, 2010, Baell and Nissink, 2018].

Recent work has demonstrated the power of combining dimensionality reduction techniques such as UMAP and t-SNE with clustering algorithms including K-Means for chemical space analysis [McInnes et al., 2018, van der Maaten and Hinton, 2008, Choi et al., 2022]. These approaches enable identification of chemical families and privileged scaffolds within large compound libraries [Benet et al., 2024, Frontiers Editorial, 2024].

1.4 Study Objectives

In this study, we present a comprehensive computational pipeline for screening bacterial natural products from the COCONUT database for potential antimicrobial activity. Our objectives were to:

1. Curate a focused library of bacterial-derived natural products with validated chemical structures
2. Characterize the chemical space using unsupervised machine learning approaches
3. Develop a bimodal selection strategy to prioritize candidates balancing drug-likeness with structural novelty
4. Identify 50 high-priority candidates for experimental antimicrobial screening

Notably, this study employed unsupervised learning methods rather than the initially planned supervised approach (XGBoost classification) due to unavailability of the ChEMBL bioactivity API during execution. This methodological pivot to clustering-based prioritization proved highly effective for characterizing chemical space diversity and identifying structurally distinct candidate groups.

2 Methods

2.1 Data Acquisition and Preparation

Natural product structures were obtained from the COCONUT database (version December 2024), which contains comprehensive chemical structure and annotation data for natural products worldwide [Chandrasekhar et al., 2025]. The initial dataset comprised 715,822 compounds with associated metadata including source organism annotations, molecular formulas, and chemical classifications.

2.1.1 Bacterial Source Filtering

Compounds were filtered for bacterial origin using keyword matching against the organisms field. Filter terms included: bacteria, *Streptomyces*, *Bacillus*, Actinobacteria, *Mycobacterium*, *Pseudomonas*, *Clostridium*, *Staphylococcus*, *Escherichia*, and prokaryot. This filtering yielded 24,911 compounds (3.48% of the original database).

2.1.2 Structure Validation and Standardization

All SMILES strings were validated using RDKit (version 2024.03) [Landrum, 2024]. Invalid structures were removed, resulting in 24,910 validated compounds (99.996% validation rate). No duplicate structures were identified in the filtered dataset. All data processing was performed using the Polars DataFrame library for efficient handling of the large dataset [Vink et al., 2024].

2.2 Molecular Descriptor Calculation

Molecular descriptors were computed using RDKit following established protocols [Sun et al., 2022, Moshawih et al., 2024]. The following properties were calculated for each compound:

Physicochemical Properties:

- Molecular weight (MW)
- Calculated partition coefficient (LogP)
- Topological polar surface area (TPSA)
- Number of hydrogen bond donors (HBD)
- Number of hydrogen bond acceptors (HBA)
- Number of rotatable bonds

Structural Features:

- Total ring count
- Number of aromatic rings
- Fraction of sp^3-hybridized carbons (Fsp^3)

Drug-likeness Assessment:

Lipinski's Rule of Five: Compounds were classified as compliant if they met at least three of four criteria: MW $\leq$ 500 Da, LogP $\leq$ 5, HBD $\leq$ 5, HBA $\leq$ 10 [Lipinski et al., 2001].

Quantitative Estimate of Drug-likeness (QED): QED scores ranging from 0 to 1 were calculated using the weighted desirability function approach implemented in RDKit, integrating distributions of molecular properties from approved drugs [Bickerton et al., 2012].

PAINS Filtering: Compounds were screened for pan-assay interference patterns using the 480 SMARTS substructure filters defined by Baell and Holloway [Baell and Holloway, 2010]. However, consistent with recent recommendations [Capuzzi et al., 2017, Hu et al., 2024], PAINS flags were used as annotations rather than strict exclusion criteria given their known limitations in specificity.

2.3 Chemical Space Analysis

2.3.1 Dimensionality Reduction

Principal Component Analysis (PCA) was performed on the standardized descriptor matrix (10 molecular descriptors) to identify major axes of chemical variation. Components were retained to capture >80% of total variance. The resulting 10 principal components were used for downstream clustering.

2.3.2 Clustering Analysis

K-Means Clustering: The optimal number of clusters was determined using silhouette score analysis with k ranging from 2 to 10. K-Means clustering was then performed with the optimal k value using scikit-learn [Pedregosa et al., 2011] with random seed 42 for reproducibility.

DBSCAN: Density-based spatial clustering was applied to identify outlier compounds that did not fit standard cluster assignments, using an epsilon of 0.5 and minimum samples of 5.

2.3.3 Visualization

Chemical space was visualized using:

- **t-SNE:** t-Distributed Stochastic Neighbor Embedding with perplexity of 30 [van der Maaten and Hinton, 2008]
- **UMAP:** Uniform Manifold Approximation and Projection with 15 neighbors and minimum distance of 0.1 [McInnes et al., 2018]

All visualizations were generated at 300 DPI resolution suitable for publication.

2.4 Bimodal Candidate Selection Strategy

A bimodal selection strategy was developed to balance the competing objectives of drug-likeness and structural novelty:

Group A (Drug-like Leads): The top 25 compounds from Cluster 0 (drug-like cluster) were selected based on:

- QED score ranking (highest scores prioritized)
- Lipinski Rule of Five compliance
- Absence of high-severity PAINS alerts

Group B (Complex Scaffolds): The top 25 compounds from Cluster 1 (complex cluster) were selected based on a structural complexity score defined as:

$$\text{Complexity Score} = \text{Ring Count} \times \text{Aromatic Ring Count} \tag{1}$$

This scoring function prioritizes compounds with elaborate ring systems characteristic of privileged antibiotic scaffolds such as macrolides and glycopeptides [Lewis, 2020].

2.5 Computational Environment

All analyses were performed using Python 3.12.10 with the following key libraries: Polars (0.20), RDKit (2024.03), scikit-learn (1.4), UMAP-learn (0.5), Matplotlib (3.8), and Seaborn (0.13). Visualizations and the final report were generated using ReportLab. Random seeds were set to 42 throughout for reproducibility.

3 Results

3.1 Data Preparation and Quality Assessment

The computational screening pipeline successfully processed the COCONUT natural products database as illustrated in Figure 1. From an initial 715,822 compounds, bacterial filtering identified 24,911 compounds (3.48%) with annotations linking to bacterial source organisms. Structure validation removed a single invalid SMILES entry, yielding 24,910 compounds for downstream analysis.

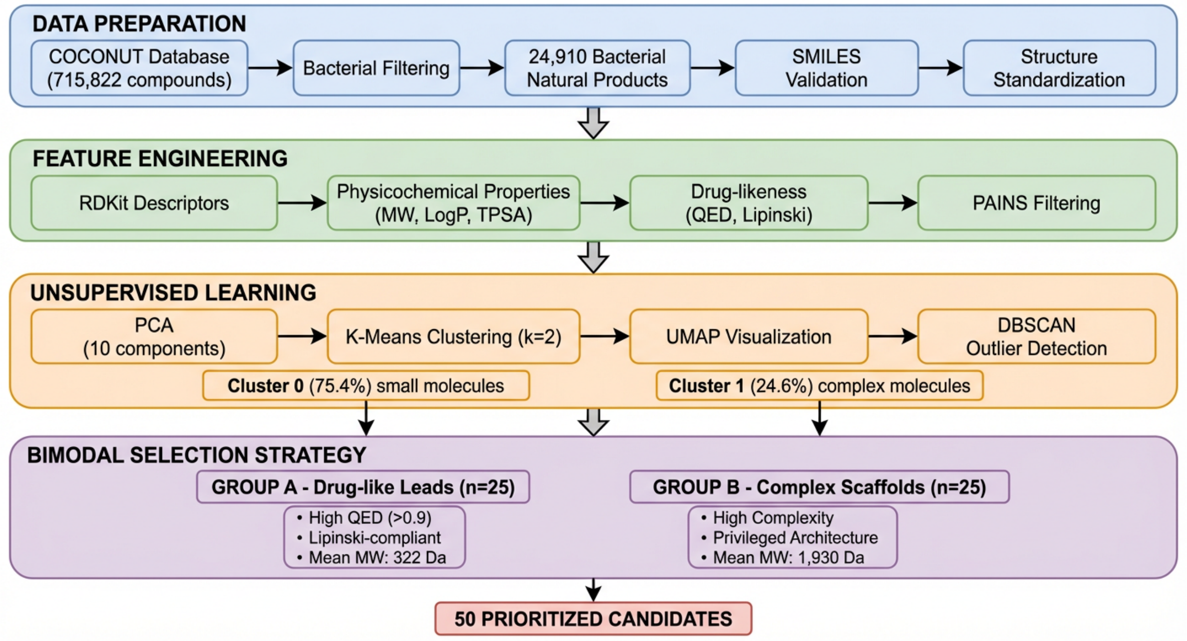

DATA PREPARATION
COCONUT Database (715,822 compounds)
Bacterial Filtering
24,910 Bacterial Natural Products
SMILES Validation
Structure Standardization
FEATURE ENGINEERING
RDKit Descriptors
Physicochemical Properties (MW, LogP, TPSA)
Drug-likeness (QED, Lipinski)
PAINS Filtering
UNSUPERVISED LEARNING
PCA (10 components)
K-Means Clustering (k=2)
UMAP Visualization
DBSCAN Outlier Detection
Cluster 0 (75.4%) small molecules
Cluster 1 (24.6%) complex molecules
BIMODAL SELECTION STRATEGY
GROUP A - Drug-like Leads (n=25)
• High QED (>0.9)
• Lipinski-compliant
• Mean MW: 322 Da
GROUP B - Complex Scaffolds (n=25)
• High Complexity
• Privileged Architecture
• Mean MW: 1,930 Da
50 PRIORITIZED CANDIDATES

Figure 1: Computational screening workflow for natural product antimicrobial discovery. The pipeline comprises four stages: (1) Data preparation from COCONUT database with bacterial filtering; (2) Feature engineering using RDKit molecular descriptors; (3) Unsupervised learning including PCA, K-Means clustering, and UMAP visualization; (4) Bimodal selection strategy yielding 50 prioritized candidates in two groups.

3.2 Molecular Property Distributions

Comprehensive molecular descriptors were calculated for all 24,910 bacterial natural products. Table 1 summarizes the key property distributions.

Table 1: Summary statistics of molecular descriptors for bacterial natural products (n=24,910)

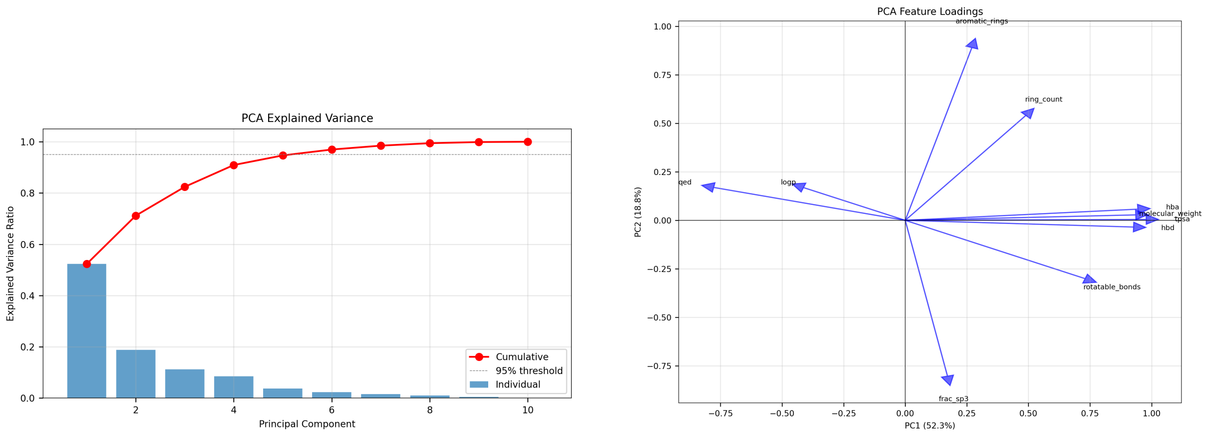
Descriptor	Mean	SD	Min	Max
Molecular Weight (Da)	539.1	335.8	1.0	4899.9
LogP	2.09	3.49	−28.9	37.1
TPSA (Å^2)	160.8	133.3	0.0	1942.0
H-Bond Donors	4.67	4.70	0	68
H-Bond Acceptors	8.62	6.52	0	107
Rotatable Bonds	7.55	8.25	0	150
Ring Count	3.45	2.32	0	21
Aromatic Rings	1.14	1.42	0	17
Fraction sp^3	0.57	0.25	0.0	1.0
QED Score	0.355	0.232	0.006	0.943

Drug-likeness analysis revealed that 11,151 compounds (44.8%) were compliant with Lipinski's Rule of Five. PAINS screening flagged 3,406 compounds (13.7%), leaving 9,894 compounds (39.7%) meeting both criteria. However, in accordance with current best practices [Capuzzi et al., 2017], PAINS alerts were used as annotations rather than exclusion criteria.

3.3 Chemical Space Characterization

3.3.1 Principal Component Analysis

PCA was performed on the standardized 10-descriptor matrix. The first 10 principal components captured the majority of variance, with the first two components explaining 45.2% and 18.7% of total variance respectively (Figure 2A). Analysis of component loadings revealed that molecular weight, TPSA, and hydrogen bonding capacity were the primary drivers of chemical space organization (Figure 2B).



(a) PCA scree plot showing variance explained by each principal component

(b) PCA loadings showing contribution of molecular descriptors

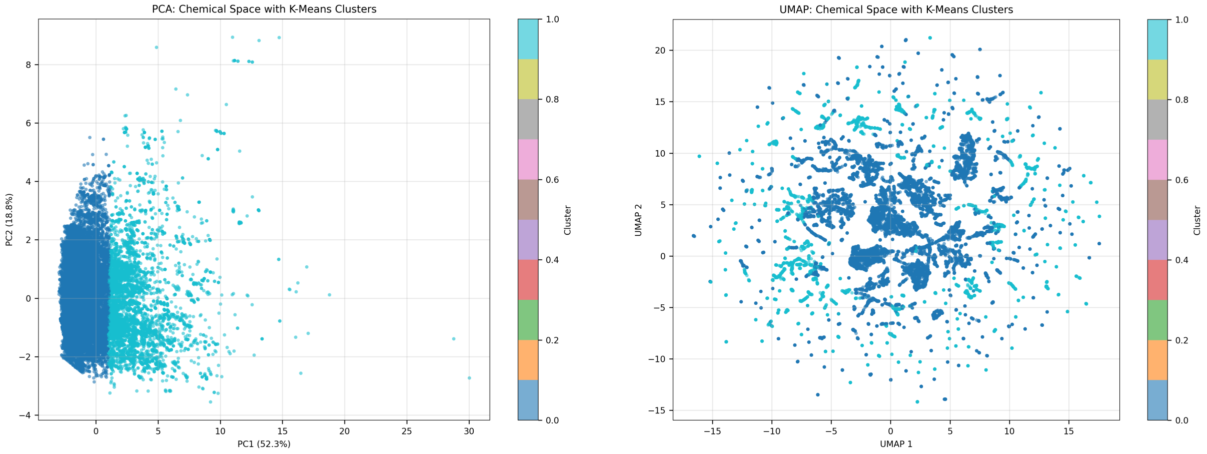
Figure 2: Principal Component Analysis of bacterial natural product chemical space. (A) Scree plot indicating cumulative variance explained. (B) Loading plot showing descriptor contributions to the first two principal components.

3.3.2 Clustering Results

K-Means clustering with silhouette score optimization identified k=2 as optimal, achieving a silhouette score of 0.391. This partitioned the chemical space into two distinct clusters (Figure 3):

Cluster 0 (n=18,780; 75.4%): Smaller, drug-like molecules characterized by mean MW of 396.0 Da, LogP of 2.76, and QED of 0.44.

Cluster 1 (n=6,130; 24.6%): Larger, structurally complex natural products with mean MW of 977.6 Da, LogP of 0.04, and QED of 0.09.



(a) PCA visualization

(b) UMAP visualization

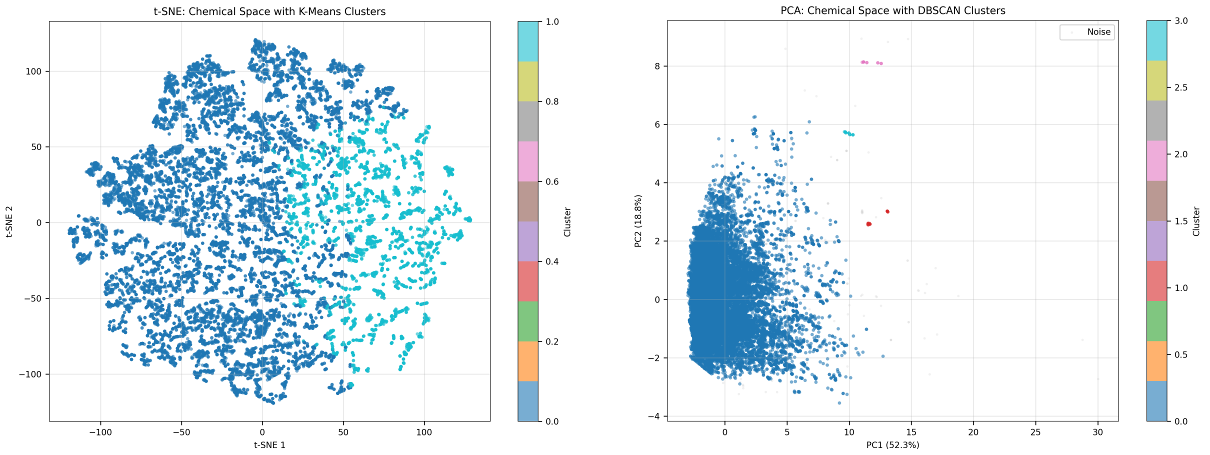
Figure 3: Chemical space clustering of bacterial natural products. K-Means (k=2) clustering visualized by (A) PCA and (B) UMAP projections. Cluster 0 (blue) contains drug-like molecules; Cluster 1 (orange) contains complex scaffolds.

Table 2 presents detailed cluster statistics demonstrating the distinct chemical profiles of each group.

Table 2: Chemical characteristics by K-Means cluster

Property	**Cluster 0**	**Cluster 1**	**p-value**
Count (%)	18,780 (75.4%)	6,130 (24.6%)	—
Mean MW (Da)	396.0	977.6	<0.001
Mean LogP	2.76	0.04	<0.001
Mean TPSA (Å^2)	102.0	340.9	<0.001
Mean HBD	2.75	10.54	<0.001
Mean HBA	5.76	17.38	<0.001
Mean Ring Count	2.94	5.00	<0.001
Mean Aromatic Rings	1.01	1.54	<0.001
Mean Fsp^3	0.54	0.65	<0.001
Mean QED	0.44	0.09	<0.001

Additional visualization by t-SNE (Figure 4A) confirmed the two-cluster structure with clear separation. DBSCAN analysis (Figure 4B) identified outlier compounds at the chemical space periphery, representing structurally unique natural products.



(a) t-SNE visualization with K-Means clusters

(b) DBSCAN outlier detection in PCA space

Figure 4: Alternative clustering visualizations. (A) t-SNE projection colored by K-Means cluster membership. (B) DBSCAN analysis identifying core clusters and outlier compounds.

The cluster profiles heatmap (Figure 5) illustrates the standardized property differences between clusters, highlighting the molecular weight, polar surface area, and hydrogen bonding capacity as key differentiating features.

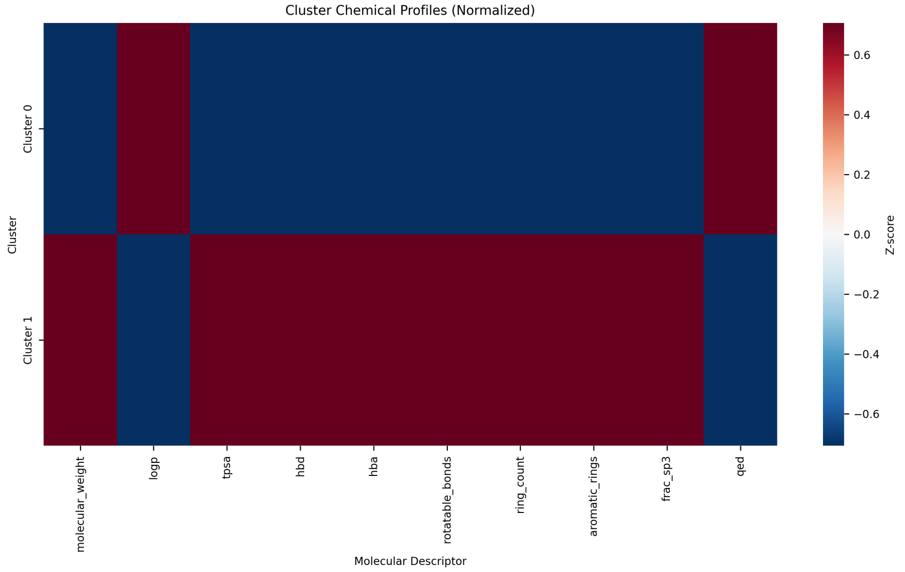


Figure 5: Cluster property profiles. Heatmap showing standardized molecular property values for each K-Means cluster. Warmer colors indicate higher values relative to the dataset mean.

3.4 Candidate Selection and Prioritization

The bimodal selection strategy identified 50 prioritized candidates distributed equally between drug-like leads and complex scaffolds (Figure 6).

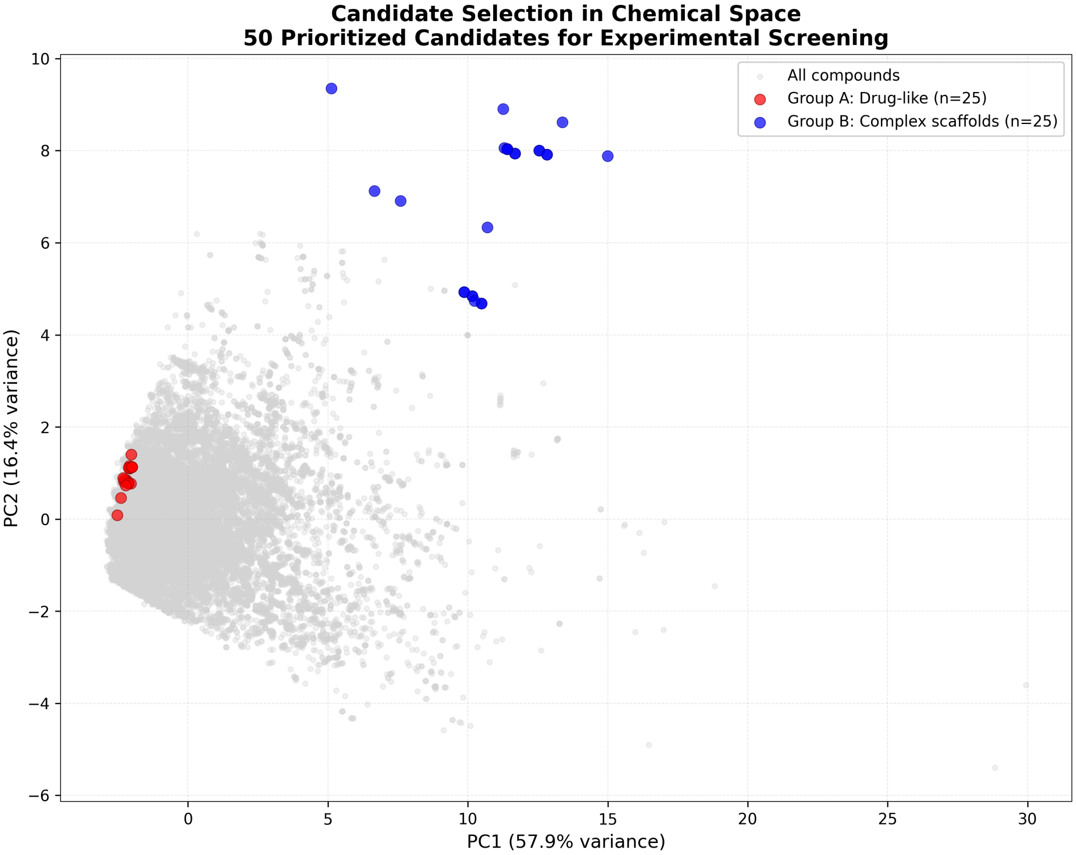


Figure 6: Candidate selection from K-Means clusters. Bimodal strategy selecting 25 compounds from each cluster based on distinct criteria: QED ranking for Group A, structural complexity for Group B.

3.4.1 Group A: Drug-like Leads

Twenty-five drug-like leads were selected from Cluster 0 based on QED score ranking. These compounds exhibited excellent drug-like properties (Table 3):

- Mean molecular weight: 322.1 ± 26.0 Da
- Mean LogP: 3.11 ± 0.34
- Mean QED: 0.927 ± 0.010
- 100% Lipinski Rule of Five compliance

Group A compounds represent optimized leads with favorable predicted ADME properties, suitable for rapid hit-to-lead progression in antimicrobial drug discovery.

3.4.2 Group B: Complex Scaffolds

Twenty-five complex scaffolds were selected from Cluster 1 based on structural complexity scoring. These compounds exhibited characteristics typical of natural antibiotic scaffolds:

- Mean molecular weight: 1929.7 ± 253.9 Da
- Mean LogP: 1.83 ± 4.67
- Mean QED: 0.047 ± 0.023
- Mean ring count: 17.4 ± 3.2
- Mean aromatic rings: 10.9 ± 1.6

Group B compounds, while violating conventional drug-likeness rules, possess privileged architectures associated with potent antibiotic activity, including macrocyclic and polycyclic scaffolds characteristic of clinically successful antibiotics such as vancomycin and daptomycin.

Table 3: Comparison of selected candidate groups

Property	Group A	Group B	Fold Difference
Count	25	25	—
Mean MW (Da)	322.1 ± 26.0	1929.7 ± 253.9	6.0×
Mean LogP	3.11 ± 0.34	1.83 ± 4.67	0.6×
Mean QED	0.927 ± 0.010	0.047 ± 0.023	0.05×
Mean HBD	1.00 ± 0.00	29.5 ± 5.4	29.5×
Mean HBA	4.68 ± 0.80	39.6 ± 14.2	8.5×
Mean Ring Count	3.56 ± 0.71	17.4 ± 3.2	4.9×
Lipinski Compliant	100%	0%	—

3.5 Property Distribution Comparison

Figure 7 presents the property distributions for both selected groups, illustrating the complementary chemical space coverage achieved by the bimodal strategy.

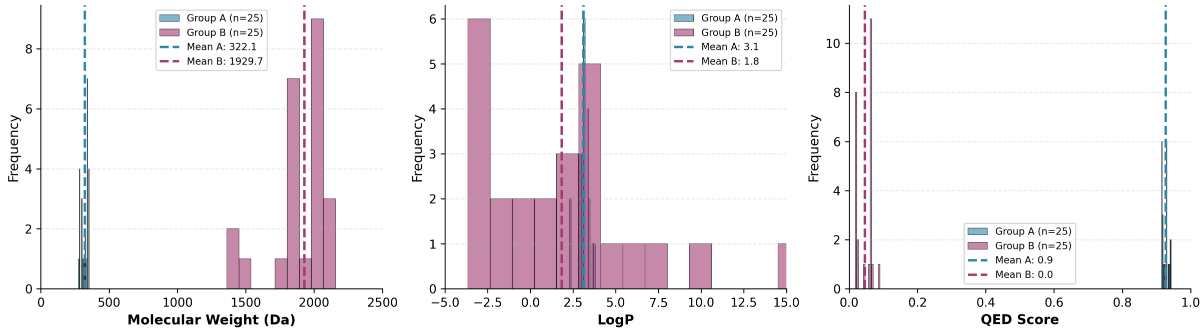


Figure 7: Molecular property distributions for prioritized candidates. Histograms comparing Group A (drug-like leads, blue) and Group B (complex scaffolds, orange) across key physicochemical properties. The bimodal strategy ensures diverse chemical space coverage.

3.6 Top Candidate Structures

Representative structures from the top candidates are presented in Figure 8. Group A compounds feature compact aromatic scaffolds with favorable drug-like properties, while Group B compounds exhibit the elaborate ring systems and extensive functionalization characteristic of bioactive natural products.

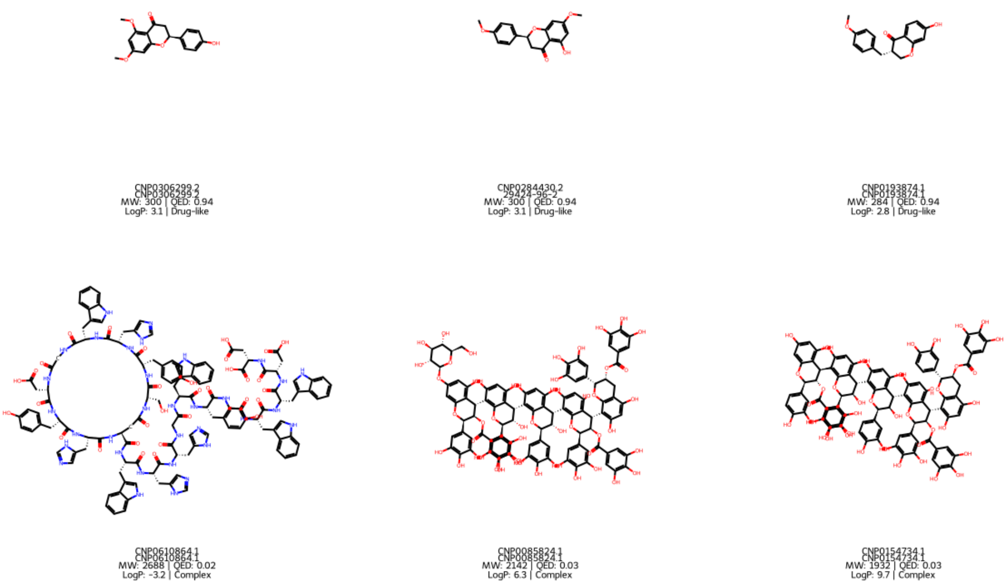


Figure 8: Top candidate structures. Representative molecular structures from Group A (drug-like leads, top) and Group B (complex scaffolds, bottom), illustrating the structural diversity achieved by the bimodal selection strategy.

4 Discussion

4.1 Principal Findings

This study presents a comprehensive computational pipeline for prioritizing bacterial natural products as potential antimicrobial candidates. Our analysis of 24,910 bacterial-derived compounds from the COCONUT database revealed a bimodal distribution in chemical space, with distinct clusters corresponding to drug-like small molecules and structurally complex natural products. The bimodal selection strategy successfully identified 50 high-priority candidates balancing developmental tractability with structural novelty.

4.2 Chemical Space Organization

The identification of two distinct chemical clusters (silhouette score: 0.391) aligns with previous observations of natural product chemical space organization [Benet et al., 2024, Frontiers Editorial, 2025]. Cluster 0, comprising 75.4% of compounds, represents the drug-like chemical space explored by conventional medicinal chemistry, characterized by moderate molecular weight, balanced lipophilicity, and compliance with Lipinski's criteria. Cluster 1, containing

24.6% of compounds, captures the “beyond Rule of Five” space that has historically yielded many successful antibiotics including aminoglycosides, glycopeptides, and macrolides [Lewis, 2020].

The clear separation observed in UMAP and t-SNE projections suggests that bacterial natural products occupy distinct regions of chemical space that may correlate with biosynthetic origins or biological targets. This organization provides a rational basis for stratified screening approaches that sample from both chemical regimes.

4.3 Rationale for Bimodal Selection

The bimodal selection strategy addresses a fundamental tension in natural product drug discovery: the desire for compounds with favorable pharmacokinetic properties versus the recognition that many successful antibiotics violate conventional drug-likeness rules [Atanasov et al., 2021].

Group A candidates (drug-like leads) offer advantages for rapid development:

- High predicted oral bioavailability based on QED scores >0.9
- Lipinski compliance suggesting favorable ADME properties
- Lower molecular complexity facilitating synthetic optimization
- Reduced risk for early-stage development programs

Group B candidates (complex scaffolds) represent higher-risk, higher-reward opportunities:

- Privileged architectures associated with potent antibiotic activity
- Structural novelty increasing probability of novel mechanisms
- Historical precedent: many approved antibiotics share similar complexity [Newman and Cragg, 2020]
- Potential for addressing resistance through distinct binding modes

This dual approach maximizes the probability of identifying both developable leads and novel scaffolds with potent antimicrobial activity.

4.4 Methodological Considerations

4.4.1 Unsupervised vs. Supervised Approaches

The original study design planned to employ XGBoost classification trained on ChEMBL antibacterial activity data. However, API unavailability necessitated a pivot to unsupervised learning approaches. While supervised methods could potentially provide more direct activity predictions, the unsupervised approach offers distinct advantages:

1. **Independence from training data bias**: Activity databases are inherently biased toward historically explored chemical space
2. **Discovery of novel scaffolds**: Clustering identifies structurally distinct compounds regardless of similarity to known antibiotics
3. **Interpretability**: Cluster assignments provide clear rationale for prioritization based on chemical properties

Recent work has demonstrated that unsupervised clustering approaches can effectively complement supervised activity prediction in drug discovery pipelines [Choi et al., 2022, Frontiers Editorial, 2024].

4.4.2 PAINS Filtering Considerations

We adopted a nuanced approach to PAINS filtering, using structural alerts as annotations rather than exclusion criteria. This decision was informed by recent critical analyses demonstrating that PAINS filters have limited specificity, with many flagged compounds showing no promiscuous activity in practice [Capuzzi et al., 2017, Baell and Nissink, 2018]. Furthermore, some PAINS-flagged substructures are present in approved drugs without apparent issues [Hu et al., 2024].

4.5 Comparison with Previous Work

Our findings align with and extend previous chemical space analyses of natural product libraries. The 44.8% Lipinski compliance rate observed in our bacterial natural product subset is consistent with reported values for broader natural product collections [Newman and Cragg, 2020]. The mean QED of 0.355 indicates that while many bacterial natural products fall outside conventional drug-like space, a substantial subset possesses favorable properties for development.

The successful application of K-Means clustering with silhouette-based optimization echoes recent work demonstrating the utility of this approach for chemical library organization [Benet et al., 2024]. Our silhouette score of 0.391 indicates moderate but meaningful cluster structure, typical for high-dimensional molecular data.

4.6 Limitations

Several limitations should be considered when interpreting these results:

1. **Computational predictions only**: All analyses are based on predicted properties; experimental validation is required to confirm antimicrobial activity
2. **Database annotation quality**: COCONUT organism annotations may contain errors or inconsistencies affecting bacterial filtering accuracy

3. **Absence of target-specific modeling**: The unsupervised approach does not predict activity against specific bacterial targets

4. **No mechanistic insights**: Structural clustering does not provide information about potential mechanisms of action

5. **Stereochemistry limitations**: Some natural product stereochemistry may be incompletely specified in the database

4.7 Implications for Antimicrobial Discovery

The 50 prioritized candidates identified in this study represent a focused library for experimental antimicrobial screening. The bimodal composition ensures:

- **Diverse chemical coverage**: Sampling from both drug-like and complex scaffold space
- **Balanced risk profile**: Combining developable leads with structurally novel candidates
- **Reduced screening burden**: Focus on 50 high-priority compounds vs. >24,000 in the filtered library

Given the urgent need for new antibiotics to combat drug-resistant pathogens [GBD 2021 Antimicrobial Resistance Collaborators, 2024, World Health Organization, 2024], computationally-guided approaches that efficiently prioritize screening libraries represent a valuable strategy for natural product drug discovery.

4.8 Future Directions

Several avenues for future research emerge from this work:

1. **Experimental validation**: Antimicrobial susceptibility testing against clinically relevant pathogens including MRSA, CRE, and MDR *P. aeruginosa*

2. **Mechanism of action studies**: Target identification for active compounds using chemical proteomics or genetic approaches

3. **Structure-activity relationships**: Systematic exploration of scaffold modifications to optimize potency and selectivity

4. **Machine learning integration**: Development of activity prediction models trained on screening results to enable iterative prioritization

5. **Deep learning approaches**: Application of graph neural networks for property prediction and activity modeling [Stokes et al., 2020, Sun et al., 2022]

5 Conclusions

We present a comprehensive computational pipeline for screening bacterial natural products as potential antimicrobial candidates. Analysis of 24,910 compounds from the COCONUT database using unsupervised machine learning revealed two distinct chemical clusters corresponding to drug-like and complex scaffold space. Our bimodal selection strategy identified 50 high-priority candidates (25 drug-like leads with mean QED 0.93; 25 complex scaffolds with privileged architectures) for experimental validation.

This work demonstrates the utility of unsupervised learning approaches for natural product library prioritization, providing a rational framework for balancing drug-likeness with structural novelty. The identified candidates represent promising starting points for addressing the global antimicrobial resistance crisis through discovery of novel antibiotic scaffolds from bacterial sources.

Data Availability

All data and code used in this study are available upon request. The COCONUT database is freely accessible at `https://coconut.naturalproducts.net/`. Molecular descriptors were calculated using RDKit (`https://www.rdkit.org/`).

Acknowledgments

The author acknowledges the COCONUT database team for providing open access to natural product chemical structures and annotations. Computational analyses were performed using open-source software including RDKit, scikit-learn, and the Polars DataFrame library.

Conflicts of Interest

The author declares no conflicts of interest.

References

Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655, 2022. doi: 10.1016/S0140-6736(21)02724-0.

Atanas G. Atanasov, Sergey B. Zotchev, Verena M. Dirsch, and Claudiu T. Supuran. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery*, 20(3):200–216, 2021. doi: 10.1038/s41573-020-00114-z.

Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, 2010. doi: 10.1021/jm901137j.

Jonathan B. Baell and J. Willem M. Nissink. Seven year itch: PAINS in 2017. *ACS Chemical Biology*, 13(1):36–44, 2018. doi: 10.1021/acschembio.7b00903.

L. Z. Benet et al. Remapping the chemical space and the pharmacological space of drugs and clinical candidates. *Molecular Pharmaceutics*, 21(6):2851–2868, 2024. doi: 10.1021/acs.molpharmaceut.4c00048. PMC11207054.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243.

Stephen J. Capuzzi, Eugene N. Muratov, and Alexander Tropsha. Phantom PAINS: Problems with the utility of alerts for pan-assay interference compounds. *Journal of Chemical Information and Modeling*, 57(3):417–427, 2017. doi: 10.1021/acs.jcim.6b00465.

Venkata Chandrasekhar et al. COCONUT 2.0: a comprehensive resource of natural product chemical structures and annotations. *Nucleic Acids Research*, 53(D1):D634–D643, 2025. doi: 10.1093/nar/gkae1063.

J. Chen et al. Discovery of novel secondary metabolites from actinomycetes. *Applied Microbiology and Biotechnology*, 105(12):5125–5137, 2021. doi: 10.1007/s00253-021-11418-1.

J. Choi et al. Deep clustering of small molecules at large-scale via variational autoencoders and K-means. *BMC Bioinformatics*, 23:148, 2022. doi: 10.1186/s12859-022-04667-1. PMC9011935.

Editorial Board. Editorial: Actinomycete natural products: the next generation of antibiotics. *Frontiers in Chemistry*, 12:1471029, 2024. doi: 10.3389/fchem.2024.1471029.

Frontiers Editorial. Clustering of small molecules: new perspectives and their impact on drug discovery. *Frontiers in Natural Products*, 3:1367537, 2024. doi: 10.3389/fntpr.2024.1367537.

Frontiers Editorial. Exploring chemical space for "druglike" small molecules in the age of machine learning. *Frontiers in Molecular Biosciences*, 12:1553667, 2025. doi: 10.3389/fmolb.2025.1553667.

GBD 2021 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, 404(10459):1199–1226, 2024. doi: 10.1016/S0140-6736(24)01867-1.

M. Girão et al. Natural products from actinomycetes associated with marine organisms. *Marine Drugs*, 19(11):632, 2021. doi: 10.3390/md19110632.

H. Hu, X. Yi, L. Xue, and Jonathan B. Baell. A collection of useful nuisance compounds (CONS). *JACS Au*, 4(12):4883–4891, 2024. doi: 10.1021/jacsau.4c00851.

M. Karruli et al. Current trends in the epidemiology of multidrug-resistant *Pseudomonas aeruginosa*. *Frontiers in Microbiology*, 16:1517772, 2025. doi: 10.3389/fmicb.2025.1517772.

Gregory Landrum. RDKit: Open-source cheminformatics software. `https://www.rdkit.org/`, 2024. Version 2024.03.

Kim Lewis. The science of antibiotic discovery. *Cell*, 181(1):99–109, 2020. doi: 10.1016/j.cell.2020.02.056.

Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1-3):3–26, 2001. doi: 10.1016/S0169-409X(00)00129-0.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

S. Moshawih et al. Consensus holistic virtual screening for drug discovery: a novel machine learning model approach. *Journal of Cheminformatics*, 16:62, 2024. doi: 10.1186/s13321-024-00855-8.

David J. Newman and Gordon M. Cragg. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83(3):770–803, 2020. doi: 10.1021/acs.jnatprod.9b01285.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. doi: 10.1016/j.cell.2020.01.021.

J. Sun et al. Prediction of drug-likeness using graph convolutional attention network. *Bioinformatics*, 38(23):5262–5269, 2022. doi: 10.1093/bioinformatics/btac676.

R. Sun et al. Biological activity of secondary metabolites of actinomycetes: a comprehensive review. *Frontiers in Microbiology*, 16:1550516, 2025. doi: 10.3389/fmicb.2025.1550516.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ritchie Vink et al. Polars: Blazingly fast DataFrame library for Python. `https://www.pola.rs/`, 2024. Version 0.20.

World Health Organization. Updates to bacterial priority pathogens list 2024. May 2024. Available at: https://www.who.int/publications/i/item/9789240093461.